DOCUMENT RESUME

ED 377 711 FL 022 686

AUTHOR Jones, Daniel; Alexa, Melina
TITLE Towards Automatically Aligning German Compounds with English Word Groups in an Example-Based Translation System.
PUB DATE [94]
CONTRACT GR/G43546
NOTE 6p.; This research was sponsored by the Science and Engineering Research Council.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *English; Foreign Countries; *German; *Language Patterns; *Language Processing; Lexicology; *Machine Translation; *Vocabulary
IDENTIFIERS *Compound Words

ABSTRACT

As part of the development of a completely sub-symbolic machine translation system, a method for automatically identifying German compounds was developed. Given a parallel bilingual corpus, German compounds are identified along with their English word groupings by statistical processing alone. The underlying principles and the design process are described here. Design began with small-scale word-alignment experiments, using 2,543 English words and 1,898 German words that yielded unique lexical items in each language. A technique for decreasing reliance on one-to-one word correspondences was then applied, resulting in a distinct ability to capture relationships between compounds and non-compounded expressions. Statistical analysis of these relationships provides data on which to base machine translation operations. It is concluded that the method used is effective on identifying cross-language lexical fertility to establish translation units for re-combination within the example-based translation process. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Daniel B. Jones
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED 377 711

Towards Automatically Aligning German Compounds with English Word Groups in an Example-based Translation System*

Daniel JONES and Melina ALEXA
Centre for Computational Linguistics
UMIST
Manchester
UK
{danny,melina}@ccl.umist.ac.uk

Abstract

An important factor in bootstrapping a completely non-symbolic Machine Translation system is an ability to identify potential translation units. A method for automatically identifying German compounds has been developed as part of an example-based MT project. Given a parallel bilingual corpus, German compounds are identified along with their English word groupings by statistical processing alone.

Keywords: Example-based MT, Corpus-based NLP, Sub-symbolic NLP, Bootstrapping.

1 Introduction

The work of various research groups over the last few years e.g. (Gale and Church, 1993; Kay and Röscheisen, 1993; Catizone et al., 1989; Brown et al., 1990) illustrates a desire to demonstrate the effectiveness of processing natural language without the need for human intervention, particularly with respect to machine translation. This is a very attractive idea as it means that much time can be saved in creating lexical and grammatical information before the system can begin translating. In contrast purely sub-symbolic processing simply entails presenting a system with sufficient bilingual material from which the system can induce the information necessary to translate source language input.

2 Motivation

The research reported here is directly motivated by a need to produce alignment data for a sub-symbolic example-based machine translation system. The MEG system (Multilingual Example-based Generation) aims to allow (monolingual) users to produce translations of a sublanguage text they are composing. As the the source text is being composed in real time (and in conjunction with the translation system) and has not been written before being presented to the translation system as is usually the case, this form of translation has been referred to as "translation without a source text" (Somers et al., 1990).

MEG aims to be as sub-symbolic as possible in the way it carries out the translation process, i.e., the information it uses to produce a translation should be produced without any human intervention i.e. knowledge in the form of human-produced grammars (monolingual and transfer) and lexicons should not be used or relied upon for maintenance reasons. As much information should be extracted from the bilingual corpora as possible so that the bootstrapping process for the incorporation of new languages and the extension of translation capability should be achieved as quickly and as cleanly as possible.

The system architecture (at its simplest level) incorporates processes for matching input text with translation examples in the bilingual corpus and a recombination process which recombines elements of high-scoring matches extracted from the corpus. As the recombination process deals with source and target language translation fragments of various sizes, the re-combinator will be greatly facilitated if it has some confidence in how the translation units align with each other. The motivation for producing static (produced off-line) and dynamic (produced at run-time) alignment data is then clear.

3 Methodology

The bilingual corpus used in the experiments was a extended manually sententiallyaligned English-German version of the Avalanche Bulletin material obtained from ISSCO in Geneva (Bouillon et al., 1992). The exploratory initial experiments were per-

This research was sponsored by the Science and Engineering Research Council. Grant number GR/G43546.

FL022686

BEST COPY AVAILABLE

formed on a small subset of this English-German material. As the requirement for sentential alignment had already been satisfied there was no need to align the using sentences by the techniques described by Gale and Church (ibid.).

Kay and Röscheisen (ibid.) have demonstrated interesting results in the area of word alignment. The results they report for word alignment, i.e., the probability that source word S is a translation of target word T, appear encouraging. From their first fifty word alignments they illustrate they claim (approximately) a 84% success rate.

The procedures involved in this form of word alignment can be expressed generally as follows:

1. Make word lists for each of the separate language corpora.

2. Pair two words from the source and target language lists – as this is a statistical procedure no *a priori* judgements are made regarding which source words align with which target words so this pairing is random.

3. Locate sentences in source and target corpora which contain these words.

4. Seek justification for the current word pairing being an actual alignment. The more times this pair of words occurs in aligned sentences the greater the probability they are translations of one another. Also, the frequency each word in the pair occurs across the corpus as a whole adds weight to the probability of alignment. E.g. if a word only occurs once throughout the entire source corpus, the possibility of it producing spurious data is greater.

4 Experiments in Word Alignment

A relatively small subset of the parallel sententially-aligned English-German corpus was used for the initial word alignments experiments (approximately 40 bulletins). The raw corpus sizes were 2543 English words and 1898 German words. These corpora yielded 290 and 369 unique English and German lexical items respectively.

The algorithm described above was applied to a section of the Avalanche corpus. Additionally, as in Kay and Röscheisen's work Dice's coefficient, (van Rijsbergen, 1979) was used in order to calculate the alignment probabilities i.e.

$$\frac{2c}{N_S(s) + N_T(t)}$$

where c denotes the number of pairs of sentences in which the proposed aligned words occur. $N_T(x)$ the number of times word x word occurs in text T. For example if the words*the* and *das* occur in a text 23 and 27 times respectively, and they both occur in the same sentences 50 times, the probability that they are in alignment is 1. This seems logical as this distributional behaviour suggests that whenever these two words appear in the text they always occur in the same sentence. By the same token if a source and target word are paired but they never occur in the same sentences, it would be unreasonable to predict that they are in any kind of translation relation.

Generally, the accuracy rate was 35-40% on a single-source-word to single-target-word basis. It was observed that one of the main reasons for this relatively poor performance is the fact that German compounds are used very frequently in the corpus. It is interesting though that this program matches the individual English words which constitute the translation of a particular German compound with only that compound word, providing though different probability scores:

English	*German*	*Score*
risk	Schneebrettgefahr	0.691589
of	Schneebrettgefahr	0.391892
avalanche	Schneebrettgefahr	0.638298

or in the similar case of a type of set phrase, only one word of this phrase is matched with both or all the English equivalents. So for the phrase *to be approached with caution* (*vorsictig zu beurteilen*) it gives:

English	*German*	*Score*
approached	beurteilen	0.857143
caution	beurteilen	0.857143

In addition, it is not only German but English compounds which are problematic, as is the case with *as well as*:

English	*German*	*Score*
well	sowie	0.926829

The problem of compounding with respect to this approach is clear. The "fertility" (Brown et al., 1990) of words in an aligned English-German sentence containing compounds is much greater than sentences where no compounding occurs.

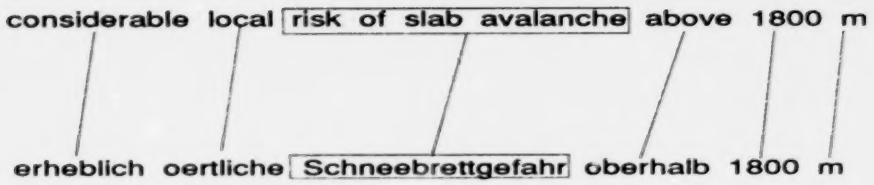

Figure 1: The German-English Compound Alignment Problem

Figure 1 on the facing page, for example, shows that the single German lexical item *Schneebrettgefahr* actually aligns with four individual English lexical items. The system as it stands will never be able to align compounds with their translation equivalents unless the compounds are either identified or a different method for hypothesising word groupings used.

Kay and Röscheisen recognise this problem:

> "The present algorithm rests on being able to identify one-to-one associations between certain words. It is clear ... that very few correspondences are noticed among everyday words ... The most interesting further developments would be in the direction of loosening up this dependence on one-to-one associations ... There are two obvious kinds of looser associations ... One would consist of connections between a single vocabulary item in one language and two or more in the other, or even between several items in one language and several in the other." (op. cit. p. 141).

5 *N*-gram Associations

Loosening up the dependence on one-to-one word relations can be done by increasing the size of the "words" used in the alignment processing. Rather then define a word as a set of characters, a word can be regarded as an *n*-gram.

For these experiments, word lists containing "words" of varying sizes were generated from the both the German and English corpora. The aim of the experiments was to establish whether trying to pair different sized "words" would capture the lexical fertility differences between the two languages.

A set of "words" for aligning was produced by moving a sliding window of varying size across the sentences of the corpus. For example, a sentence can be represented as a set of words:

$$(w_i, w_j, w_k, ..., w_N)$$

If the window is to contain three words at a time the first pass will have $i = 1, j = 2, k = 3$ and in subsequent passes, $i = i + 1$, $j = j + 1$, and so on. A range of results was generated by producing *n*-grams where *n* was as large as 7. Not surprisingly, the results showed most promise when the sizes of the source and target *n*-grams were around the same, i.e., when "word" units only differed by a maximum of three or four actual words. Alignments of 1-6 or 1-5 (source-target) were not very fruitful.

The results of these experiments showed a distinct ability to capturing the relations between compounds and non-compounded expressions. For example in an experiment where alignments were attempted between all possible linear groupings of three English words and two German words, it was discovered that German compounds and their corresponding English translations were indeed being linked.

5.1 German Compound Identification

Table 1 on the next page shows sample high-scoring alignments for three English words to two German words. Of course in this is just part of the data produced where the size of "word" used in the alignments varies from 1 to some reasonable number, e.g. 5 or 6. Such a range of data can be used to identify German compounds.

The method is quite straightforward in that for each set of data, i.e. 1-1, 2-3, 4-2, etc., the highest scoring alignment from each group is selected. These highest scoring examples are then compared. For example, if the highest scoring associations for the input German compound *Schneebrettgefahr* are extracted for each *n*-gram data set, the following associations with English word groups are revealed.

Ignoring for the moment the initial unary word value of 1 for *slab* the highest value[1] is given to the correct translation i.e. *risk of slab avalanche*. The influence of *slab*, although scoring 1, should be ig-

[1] Calculated in the same way with single word-word comparisons i.e. by using Dice's coefficient.

Three English Words	*Two German Words*	*Score*
section danger levels	Section Gefahrenstufen	1
slab avalanche above	Schneebrettgefahr ueber	0.444444
alpine slopes in	Alpennordhang im	0.769231
the rest of	im uebrigen	0.551724
steep slopes as	Steilhaengen sowie	1
drift accumulations as	Triebschneeansammlungen sowie	1
to be approached	zu beurteilen	0.857143
m there is	m besteht	0.666667
is a temporary	sind kurzfristig	1
alpine ridges northern	alpenhauptkamm noerdliches	1
of Valais and	Wallis und	1
this is restricted	diese beschraenkt	1
the entire Swiss	ganzen schweizerischen	1
swiss alpine region	schweizerischen alpengebiet	0.8
Upper Engadine including	Oberengadin inschliesslich	1
and southern Ticino	und Suedtessin	1
low avalanche risk	geringe Lawinengefahr	1
safety measures should	Sicherungsmassnahmen sollten	1
with a further	bei zusaetzlicher	1
exposed connecting roads	exponierte Verbindungswege	1
on the remaining	am uebrigen	1

Table 1: 3-2 Alignments

N-M	*Highest Scoring English Alignment*	*Score*
1-1	slab	1
2-1	moderate local	0.5
3-1	risk of slab	0.4
4-1	risk of slab avalanche	0.53
5-1	local risk of slab avalanche	0.5

Table 2: Identification of German compound *Schneebrettgefahr*

N-M	*Highest Scoring English Alignment*	*Score*
1-1	and	1
2-1	in the	0.33
3-1	are to be	0.21
4-1	risk of slab avalanche	0.15
5-1	local risk of slab avalanche	0.12

Table 3: Identification of German non-compound *und*

nored as it is subsumed by the larger n-grams. If a compound is correctly aligned with its word group translation then it makes sense that individual words from that grouping should also score highly according to Dice's coefficient as their distributional behaviour will be the same. It should be expected that the largest of the n-grams should subsume a majority of the smaller n-gram groupings. By the same token words which do not display compounding behaviour will *not* demonstrate such signs of subsumption e.g. *und*:

Note that in this case, there is only one case of subsumption i.e. 4-1/5-1 and there is a gradual reduction in the score of as the units size of the n-gram equivalents increase. This would seem to indicate that the word *und* is not part of a compound.

6 Observations and Conclusions

The results reported here go towards demonstrating the feasibility of at least alleviating the problem of German compounding when bootstrapping a statistically-based Machine Translation system. Being able to identify this type of cross-language lexical fertility is important for establishing potential trans-

lation units for the re-combinatorial stage (Jones, 1991) of the example-based translation process.

The limitations of the experiments so far have to a large extent been caused by the relatively small amount of data used. However, even when, for example, the 1-1 alignment is technically wrong, the highest alignment scored is usually contextually closely associated with what should have been the correct alignment.

Further experimentation will involve increasing the size of the bilingual data. It is assumed that results will improve as spurious distributional data will be eliminated.

References

Bouillon, P., Boesefeldt, K., and Russell, G. (1992). Compound nouns in a unification-based MT system. In *Proceedings, 3rd Conference on Applied Natural Language Processing (Trento)*, pages 209–215.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F , Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Catizone, R., Russell, G., and Warwick, S. (1989). Deriving translation data from bilingual texts. In *Proceedings of First International Acquisition Workshop*, Detroit, Michigan.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–103.

Jones, D. (1991). *The Processing of Natural Language by Analogy with Specific Reference to Machine Translation*. PhD thesis, UMIST, Manchester.

Kay, M. and Röscheisen, M. (1993). Text-translation alignment. *Computational Linguistics*, 19(1):121–143.

Somers, H., Tsujii, J., and Jones, D. (1990). Machine translation without a source text. In Karlgren, H., editor, *Proceedings of the 13th International Conference on Computational Linguistics*, volume 3, pages 271–276.

van Rijsbergen, C. (1979). *Information Retrieval*. Butterworths.